Statistical Analysis Plan
ABRA

Evaluators: University of York, Durham University
Principal investigator: Dr Kerry Bell
(maternity cover Hannah Ainsworth)

PROJECT TITLE	Small Group Reading Support (ABRA ICT & non-ICT) (Independent evaluation of small group reading support programmes to improve literacy in Year 1 pupils: a three-armed cluster randomised controlled trial)
DEVELOPER (INSTITUTION)	Coventry University, Nottingham Trent University, Concordia University
EVALUATORS (INSTITUTION)	University of York (York Trials Unit) and Durham University
PRINCIPAL INVESTIGATOR	Dr Kerry Bell (Hannah Ainsworth as maternity cover Oct 2017 – Oct 2018)
TRIAL (CHIEF) STATISTICIAN	Caroline Fairhurst
SAP AUTHORS	Caroline Fairhurst and Alex Mitchell
TRIAL REGISTRATION NUMBER	ISRCTN37208856
EVALUATION PROTOCOL URL OR HYPERLINK	Please find amended protocol here.

This statistical analysis plan was prepared following the randomisation of schools and completion of pre-tests but prior to the marking of the pre-tests and collection of outcome (post-test) data.

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	03/06/2019	Original

Differences between the protocol and SAP

- In the current version of the protocol, the two analysis models for the primary outcome are specified as follows. One model is used to obtain the comparisons between the ICT programme and control group, and between the non-ICT programme and control group. A second model, excluding any school that could not implement the ICT version of the programme and so was randomised only between the non-ICT and control groups, is used for the comparison between the ICT and non-ICT arms. However, on further reflection we feel this is incorrect and this SAP proposes the following model specifications. One model will exclude pupils in the

schools randomised to the ICT group, and will be used to investigate the difference between the non-ICT and control groups. The second model will include pupils from all three groups except those from the eight schools that were only randomised between the non-ICT and control groups (because they did not have the technology to implement the ICT programme). This model will be used to obtain the pairwise comparisons between the ICT programme and control groups, and between the ICT programme and the non-ICT programme. Two models are necessary since it is not appropriate to include schools that could never have been allocated to receive the ICT programme in a comparison involving this group.

- After attending the training, some schools allocated to the ICT or non-ICT arms felt that they were unable to deliver the programme to the number of pupils they had initially specified and allowed the Evaluation Team at York Trials Unit (YTU) to randomly select a smaller subset of their original cohort to take part in the programme, according to the number they felt they could manage. We will aim to post-test all pupils with a pre-test, but conduct a sensitivity analysis to look at the impact of this by excluding the pupils who were originally intended to receive the programme but were 'deselected' at random post-randomisation.

Table of contents

Introduction

A global aim of education is to improve standards of literacy. In England, the latest national Key Stage 2 results indicate that only 71% of pupils met the expected standard in reading (Department of Education, 2017). As such, it is important that research continues to identify effective approaches to increase literacy skills. A recent tertiary review has recommended that interventions including phonics approaches to increase reading acquisition should be evaluated in large scale RCTs (Torgerson et al., 2018). A recent EEF-funded review of the use of teaching assistants found beneficial impacts on pupil attainment when teaching assistants were used to deliver structured small group interventions (Sharples et al., 2015).

The reading support programme developed for this trial (Reading and Understanding in Key Stage 1, or RUKS) is non-targeted and takes place in Year 1 of primary school. It is a structured programme comprising 20 weeks of lesson plans involving phonics, fluency and comprehension activities. It can be delivered by school staff to small groups of Year 1 pupils, using activities via online software (Abracadabra) or with adapted, more traditional paper-based activities. The Abracadabra (ABRA) software is a freely available, computer-based, online literacy toolkit widely used in Canada (Abrami et al., 2010). ABRA provides phonics, fluency and comprehension activities around a series of age appropriate texts, and aims to increase skills in reading.

A number of small scale developer-led RCTs conducted in Canada, where the ABRA toolkit was first developed, have shown support for ABRA (Comaskey et al., 2009, Savage et al., 2009) as well as a larger effectiveness trial (Savage et al., 2013). In 2016, an EEF-funded efficacy trial of the reading support programme (RUKS) delivered online via ABRA as a computer based programme (ICT) and a paper-based programme (non-ICT) found that pupils who received the ICT or non-ICT programme were found to make between two and three months' progress in literacy compared to pupils who received standard provision. A more marked effect was observed for pupils eligible for free schools meals (FSM) and those with below average pre-test reading scores (McNally et al., 2016, Johnson et al., 2019).

Consequently, the EEF has funded an effectiveness trial to test the impact of the RUKS reading programme when delivered at scale and further investigate any differences between the ICT delivery (using ABRA) and the equivalent non-ICT, paper-based programme.

Definition of terms

ABRA = a suite of online activities to boost blending, decoding, comprehension etc.

RUKS = a 20-week programme of structured work using ABRA activities (e.g. 2 minutes of blending, 5 minutes of decoding), designed around the KS1 curriculum in Britain.

RUKS ICT = the 20 week programme delivered via the online ABRA platform.

RUKS non-ICT = the 20 week programme delivered via paper and pencil activities (adapted from the online ABRA programme).

Design overview

The current evaluation is a pragmatic three-armed cluster randomised controlled trial (RCT), with random allocation at the school level 1:1:1 to:

- ICT – schools allocated to the ICT arm will deliver the RUKS reading programme using the ABRA ICT delivery model (in addition to standard classroom phonics instruction); or

- Non-ICT – schools allocated to the non-ICT arm will deliver the RUKS reading programme using an equivalent non-ICT based delivery model (in addition to standard classroom phonics instruction); or
- Control – schools allocated to control will proceed with business as usual including any usual small group teaching.

The delivery of the ICT arm requires ready access to technology. Some schools, in particular small or rural schools, may have recurrent problems with technology. To avoid excluding schools based on insufficient ICT facilities, schools that identified potential ICT limitations were randomised only to either the non-ICT arm or the control arm, using 1:1 allocation (see Randomisation section).

RESEARCH QUESTIONS

Primary Research Questions

1. How effective is the ICT delivery model of the RUKS reading programme (ABRA), compared to the 'business as usual' group, in increasing the literacy skills of pupils in Year 1?

2. How effective is the paper-based delivery model of the RUKS reading programme, compared to the 'business as usual' group, in increasing the literacy skills of pupils in Year 1?

Secondary Research Questions

3. How effective is the ICT delivery model of the RUKS reading programme (ABRA), compared to the paper-based model, in increasing the literacy skills of pupils in Year 1?

4. How effective is the ICT delivery model of the RUKS reading programme (ABRA), compared to the 'business as usual' group, in increasing the literacy skills of pupils in Year 1 who are eligible for FSM?

5. How effective is the paper-based delivery model of the RUKS reading programme, compared to the 'business as usual' group, in increasing the literacy skills of pupils in Year 1 who are eligible for FSM?

6. How effective is the ICT delivery model of the RUKS reading programme (ABRA), compared to the paper-based model, in increasing the literacy skills of pupils in Year 1 who are eligible for FSM?

Table 1: Trial design

Trial type and number of arms	Three-armed cluster randomised controlled trial (random allocation at school level)
Unit of randomisation	School, via minimisation
Minimisation variables	• Staff type (3 levels: qualified; non-qualified; both) • Number of pupils in the Year 1 cohort (2 levels: ≤38; >38) • Percentage of pupils ever eligible for FSM in the Year 1 cohort (2 levels: ≤21%; >21%) • Geographical area (5 levels: West Midlands; East Midlands; Newcastle; Teesside; Manchester)

Primary outcome	variable	Reading ability
	measure (instrument, scale)	Progress in Reading Assessment (PiRA) test
Secondary outcome(s)	variables	• Ability to read exception, regular and nonwords • Ability to sound out single letters and letter combinations • Reading attitudes
	measures (instrument, scale)	• Diagnostic Test of Word Reading Processes (DTWRP) • Letter Sound Test (LeST) • Reading Attitudes Questionnaire (RAQ)

SCHOOLS

Schools were recruited from five recruitment 'hubs' based in the West Midlands, East Midlands, Newcastle, Teesside and Manchester. Schools with a Year 1 cohort who could feasibly deliver a reading support programme to a minimum of 10 Year 1 pupils were eligible to participate in the study. Interested schools were asked to read and agree to the requirements of participation outlined in a Memorandum of Understanding (MoU).

With the MoU, schools were asked to provide some baseline information in the School Information Sheet (SIS), including:

- how many classes and/or pupils they intended to deliver the RUKS programme to;
- the type of staff members they planned to send to the RUKS programme training;
- the school's ICT resources;
- the total number of pupils in the Year 1 cohort;
- the percentage of pupils ever eligible for FSM across the whole school cohort;
- the percentage of pupils ever eligible for FSM in the Year 1 cohort.

In addition, schools were asked to pre-identify 3-4 pupils with whom they would conduct small group teaching if they were allocated to the control group, and only if they would ordinarily undertake small group teaching (which not all schools use). Schools were asked to provide detail on the criteria they used for selecting these pupils and the small group teaching they intend to deliver.

PUPILS

The parents/carers of all pupils in the Year 1 cohorts of participating schools were sent a letter about the study. If they did not wish for their child's data to be used in the evaluation then they could return a 'Withdraw from Research Form'. These pupils would still receive the allocated programme but would not be included in outcome data collection. Furthermore, if a school felt that a particular pupil would not be suitable to receive the RUKS programme or complete the outcome measures, such pupils were excluded from the programme and evaluation.

Schools were asked to provide pupil details for all Year 1 pupils (except those for whom a withdrawal form was received) and to confirm how many Year 1 classes and/or pupils they

could feasibly deliver the programme to. Schools/pupils to take part in the evaluation were randomly selected, where possible, by YTU using the following principles:

Class selection

If the school specified the number of classes they intended to deliver the RUKS programme to:

Number of Year 1 classes in the school	*Number of classes the school intends to deliver the* RUKS *programme to*	*Selection*
1	1	None needed, all pupils in the class will be part of the evaluation (regardless of number)
>1	1 or more (*n*)	Randomly select 1 class; pre-testing of pupils in this class was 'mandatory', while for the *n*-1 other classes, testing was 'optional' i.e. preferable if the school had time and capacity. Optional testing was requested within 43 schools.

Pupil selection

If the school specified the number of pupils they intended to deliver the RUKS programme to:

Number of Year 1 classes in the school	*Number of pupils the school intends the to deliver* RUKS *programme to*	*Selection*
1	Less than the whole class	Randomly select the specified number of pupils from this class
>1	Any number (*n*)	If *n* < size of one class, then randomly select one class and then select the specified number of pupils from that class. If *n* > size of one class, then randomly select appropriate number of classes to take part to cover specified number of pupils. Classes randomly ordered, whole classes selected and then randomly selected pupils from the next class until number reached.

The majority of schools were happy for the YTU to undertake this random selection of classes/pupils to take part in the evaluation; however, a very small number of schools requested that they select the classes/pupils to take part, often for practical/logistical reasons. This was permitted as a last resort to retain the schools in the trial, and the number of schools and pupils this applies to will be reported. Since this occurred prior to randomisation, this should be balanced across the three groups so should not introduce selection bias.

Outcome measures

Primary outcome

The primary outcome measure is the Progress in Reading Assessment (PiRA) test[1], which evaluates general reading ability and in particular phonics, literal comprehension, and reading for meaning. The PiRA was used in the previous efficacy trial (McNally et al., 2016) where it was found to be a suitable outcome measure. The test takes approximately 30 minutes for a pupil to complete and is delivered in a group setting (approximately 10-15 children per group) which keeps testing costs to a minimum. At baseline (pre-test), the test was administered by school staff, but will be marked independently by research assistants employed by the evaluation team. At post-test, the PiRA will be both administered and marked independently by the appointed research assistants, employed specifically for these tasks. Test administrators and markers will be blind to allocation. Only pupils pre-tested for PiRA will be post-tested.

The PiRA Year 1 Autumn version of the test was used at pre-test, and the Summer version will be used at post-test. Both have a total raw score out of 25 obtained by summing the number of correct answers according to the established mark scheme (higher scores indicate greater attainment). From the raw score, an age-standardised score can be obtained according to the pupil's age in years and whole months (conversion tables are provided in the user manual). The advantages of using the age-standardised score rather than the raw score include:

- It is standardised to an average score of 100, immediately showing whether a pupil is above or below average, relative to PiRA's national standardisation sample;
- It allows comparisons to take into account the pupils' ages: older pupils in the year may have a higher raw scores than younger pupils, but could have a lower age-standardised score.

Therefore, the age-standardised scores will be used for analysis, as specified in the trial protocol.

Secondary outcome

All secondary outcomes will be measured post-programme only in a subset of up to 10 pupils per school randomly selected from the pupils assessed for the primary outcome at pre-test. If numbers allow, two randomly selected 'reserve' children from each school will be identified to be post-tested in the place of initially selected children who are, say, absent from school on the day of testing. All secondary outcomes will be collected/administered and marked by research assistants blind to allocation. The secondary outcomes are:

- Diagnostic Test of Word Reading Processes (DTWRP)[2]

The DTWRP assesses the reading of regular words, exception words, and non-words to enable the precise areas of difficulty experienced by individual pupils to be identified. The DTWRP takes approximately 10 minutes for a pupil to complete and is delivered on a one to one basis.

The test comprises 90 items divided as follows:

[1] More information on the PiRA can be found at **https://www.risingstars-uk.com/pira**
[2] More information on the DTWRP can be found at **https://www.gl-assessment.co.uk/products/diagnostic-test-of-word-reading-processes/**

• 30 exception words; this score provides a measure of lexical-semantic processing;
• 30 non-words; this score provides a measure of phonological recoding processing;
• 30 regular words which can be read by either process.

The DTWRP provides a pupil profile based on an overall standard age score, which will be used for analysis.

- Letter Sound Test (LeST)

The LeST assesses a person's ability to sound out single letters and letter combinations. It takes approximately 5 minutes for a pupil to complete and is delivered on a one to one basis. The number of correct items, out of 51, are summed to produce a total score. The total raw score can then be converted to an age (year group) standardised 'z-score', for 'Year 1' (ages 5-6).

- Reading Attitudes Questionnaire (RAQ)

The RAQ assesses a child's attitude and motivation in reading. It takes approximately 5 minutes for a pupil to complete and is delivered on a one to one basis. This secondary outcome aims to measure a more process-based outcome, and potentially a marker of more distal effects – since we know that there is a positive relationship between motivation and reading. It was also felt that schools would be interested in this measure.

Baseline data
Schools were asked to provide full names, unique pupil number (UPN), and date of birth (DOB) for all participating pupils at baseline. These data will allow us to request pupil-level data on Early Years Foundation Stage Profile (EYFSP) data, ever FSM status (EVERFSM_6_P), current FSM status, gender, English as an additional language and special education needs from the NPD. Schools were also asked to provide the percentage of male and female pupils, and the percentage with ever FSM status, the percentage with English as an additional language, and with Special Educational Needs, at the Year 1 cohort level (and/or participating class level).

These data will be used to describe and compare the randomised groups and in order to conduct a secondary analysis looking at the impact of the programmes on pupils with ever FSM status.

Long term follow up
Participating children may undergo standard testing at the end of Key Stage 1 (KS1; end of the 2019/2020 academic year), but it is not possible to know whether KS1 assessment will remain compulsory at that time. A further application to the NPD could be made to collect any available KS1 outcomes for participating pupils in the future. Data would likely be ready for analysis in March 2021 and consequently an addendum to the final report would be prepared after this point. Analyses for this are not specified within this SAP and would be added as an addendum should the decision to proceed with an NPD application be made in the future.

Sample size calculations overview

PROTOCOL

Overall

The previous efficacy RCT (McNally et al., 2016) found an effect size[3] of 0.138 for the ICT programme and 0.231 for the non-ICT programme, with larger effect sizes among pupils eligible for free school meals (0.368 and 0.396, respectively). A total of 84% of pupils involved at randomisation were included in the primary analysis, with an average of 40 pupils per school. The intra-cluster correlation coefficient at analysis was 0.15 and the correlation between the pre-test and the post-test was 0.43 (NB. this correlation is not the raw correlation between PiRA pre and post test - it accounts for covariates and is taken from the R-squared of a regression). The previous evaluation mandated that participating schools have a minimum of two teaching assistants and therefore the included schools are likely to be larger than the average UK primary school. Nationally, there were 27.1 pupils in the average primary school class in 2016[4].

This is a three-arm trial, with two primary research questions relating to the comparisons of the two RUKS programme arms against the shared control arm. In such a scenario, there is no consensus on whether adjustment for multiple testing is required (Wason et al., 2014). In discussion with the developer team and the EEF, the decision was made not to apply a statistical correction for the fact that we have two primary research hypotheses; therefore, both comparisons will be assessed at the 5% significance level.

We proposed to recruit a sample of 201 schools (67 in each arm). This would have given us 80% power to detect an effect size of approximately 0.20 of a standard deviation (SD) between either of the programme groups with the control group, assuming an average class size of 27, 15% attrition at the pupil-level at follow-up, an ICC of 0.15, alpha of 0.05 and a pre-post test correlation of 0.45. For the secondary outcomes, with 10 pupils per school under otherwise identical assumptions (but assuming no attrition ie actually following up 10 per school), the MDES would be approximately 0.22.

FSM

Across all primary schools in England, in January 2016, the average percentage of children claiming FSM was 14.5%[5]. In this trial, we aimed to recruit schools in deprived areas likely to have higher than average levels of pupils eligible for FSM. We will assume an average percentage of 25% in each school, this is the average observed in schools randomised into a recent EEF trial (ReflectED, still ongoing, unpublished). With an average of 27 pupils per school at randomisation, we therefore might have expected an average of 7 of them to have FSM status (201 x 27 x 0.25=1356 in total). With this number, assuming 15% pupil-level attrition at follow-up, an ICC of 0.15, alpha of 0.05 and a pre-post test correlation of 0.45, we would have 80% power to detect an effect size of 0.23 in the FSM subgroup in the primary analysis.

RANDOMISATION

Overall

[3] McNally et al. describe the calculation of the effect sizes as follows: "All the outcome variables and baseline tests have been standardised to have mean 0 and standard deviation (SD) 1 using the mean and SD of the outcomes (we have used the mean and SD for the full sample for each of the outcomes, both at post-test and at baseline respectively). This allows us to interpret the coefficients of the explanatory variables in terms of standard deviations of the outcome variable."

[4] https://fullfact.org/education/primary-class-sizes-england-and-wales/

[5] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/552342/SFR20_2016_Main_Text.pdf

In total, 166 schools were randomised into the ABRA trial. Eight of these schools specified, prior to randomisation, that they could not deliver the ICT programme; therefore, they were randomised only with the option of being randomly allocated to either the non-ICT or control arm: 3 were randomised to the non-ICT arm, and 5 to control. The remaining 158 schools did not specify that they had insufficient technology to be able to run the ICT programme. Therefore, these 158 schools were randomised to one of the three trial arms: ICT (n=51), non-ICT (n=54), and control (n=53). Overall, 51 schools were allocated to receive the ICT programme, 57 to the non-ICT group, and 58 to continue teaching as usual.

The total number of randomised pupils is defined as the number of pupils pre-tested with the PiRA (n=4015, from 157 schools). With a sample size of 4015, we would have 80% power to detect an effect size of approximately 0.22 between either of the programme groups with the control group, assuming an average of 25 pupils per school, 15% attrition at the pupil-level at follow-up, an ICC of 0.15, alpha of 0.05 and a pre-post test correlation of 0.45. For the secondary outcomes, with 10 pupils per school under otherwise identical assumptions, the MDES would be approximately 0.26.

FSM

The approximate average percentage FSM in the Year 1 cohorts of the randomised schools was 23%; therefore, we might expect 923 randomised pupils to have FSM status (approximately 6 per school). With this number, assuming 15% pupil-level attrition at follow-up, an ICC of 0.15, alpha of 0.05 and a pre-post test correlation of 0.45, we would have 80% power to detect an effect size of approximately 0.28 in the FSM subgroup for the primary analysis comparisons between the programme groups and the control group.

Table 2: Sample size scenarios for the two-arm comparison between either of the programme arms with the control arm

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
MDES[a]		0.20	0.23	0.22	0.28
Pre-test/ post-test correlations	level 1 (pupil)	0.45	0.45	0.45	0.45
	level 2 (class)	0	0	0	0
	level 3 (school)	0	0	0	0
Intracluster correlations (ICCs)	level 2 (class)	0	0	0	0
	level 3 (school)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		two-sided	two-sided	two-sided	two-sided
Average cluster size		27	7	25	6
Number of schools	ICT	67	67	47	47
	non-ICT	67	67	56	56
	control	67	67	54	54
	total	201	201	157	157

Number of pupils	ICT	1809	452	1471	338
	non-ICT	1809	452	1159	267
	control	1809	452	1385	318
	total	5427	1356	4015	923

[a] for the comparison between either of the RUKS programme arms and the control arm

Randomisation

Schools were randomised using minimisation via the software MinimPY (Saghaei and Saghaei, 2011) by statisticians from the York Trials Unit, University of York, who were not involved with the recruitment of schools. The minimisation factors were staff type, geographical area, number of pupils in the Year 1 cohort, and percentage of pupils with ever FSM status in the Year 1 cohort. Staff type refers to the member of staff identified by the school who would deliver the RUKS programme (e.g. teacher, teaching assistant, SENCO, literacy coordinator) should the school be allocated to either the ICT or non-ICT arm. The variable was categorised in the following way for the minimisation: qualified teacher(s); non-qualified teacher(s); and mix of both, using the assumption that all teachers/deputy head teachers/head teachers/SENCOs are 'qualified' and other job roles are not 'qualified' teachers. The geographical areas represented the recruitment 'hubs': West Midlands; East Midlands; Newcastle; Teesside; and Manchester. School-level data were entered into a bespoke trial database as schools returned their MoU and SIS. This data was accessed on 14/09/2018 when the minimisation program was being set up in MinimPY, for all schools who had returned data by this time. The median number of pupils and the %everFSM in the Year 1 cohort were calculated for use as cut offs in the minimisation. The median number of pupils in the Year 1 cohort was calculated from 156 schools as 38.5; therefore, the categories ≤38 and >38 were used. The median %everFSM in the Year 1 cohort was calculated from 147 schools (lower due to some with missing data at this time) as 21; therefore, the categories ≤21 and >21 were used.

Two minimisation programs were created. They had identical specifications except that one randomised the schools to one of the three groups (ICT, non-ICT, or control), and the other randomised the schools to one of only two groups (the non-ICT or control group). The second program was only used for schools that specified, prior to randomisation, that they did not have the technology to implement the ICT version of the RUKS programme.

The initial intended sequence in recruiting and randomising schools was as follows: school returns completed MoU and relevant baseline data; school eligibility checked; eligible schools asked to provide Year 1 pupil details; YTU selects classes/pupils to take part and sends schools sufficient number of PiRA pre-test papers; completed PiRA test papers returned to YTU; school randomised and informed of their allocation. Teachers from schools allocated to the ICT or non-ICT groups then attended training. Dates for the training days in the various recruitment hub locations were set in advance and schools were informed of the dates and asked to make provisions for allowing teachers to attend if necessary.

However, there was only a short timeframe at the start of the Autumn 2018 term in which to complete these pre-randomisation tasks in order to be able to inform schools of their allocation in advance of the arranged training days. Therefore, for some schools, compromises had to be made, as detailed below.

Two schools were randomised with missing data for staff type, number of pupils and/or %everFSM. To be able to include these schools in the minimisation, it was assumed that staff type was 'qualified' (2 schools), number of pupils was ≤38 (1 school), and %everFSM was ≤21 (2 schools). It has since been confirmed that for one of the schools, the %everFSM was 5% so the correct level was used. The other data are still unknown.

Due to issues with timing, and in the interest of including as many schools as possible in the evaluation, some schools had to be randomised and informed of their allocation prior to completing pre-tests, and indeed some of these schools never completed pre-tests. Of the randomised schools, 157 completed and returned their PiRA pre-tests to be marked by the evaluation team (of these, 33 schools completed the PiRA tests after being informed of their trial allocation) and one school completed the PiRA tests and claimed to have returned them via the post but they have never reached the YTU. We will only post-test pupils with a pre-test (regardless of whether this was completed before or after their school being informed of their random allocation); therefore, any school for which no valid pre-tests are returned will not continue in the evaluation. Post-testing will therefore not be completed in the school for which the pre-tests went missing in the post, but they will continue to deliver the RUKS programme. Attrition will ultimately be calculated based on the number of pupils post-tested out of those with a valid pre-test.

Additionally, after attending the training, some schools allocated to the ICT or non-ICT arms felt that they were unable to deliver the RUKS programme to the number of pupils they had initially specified and allowed the YTU to randomly select a smaller subset of their original cohort to take part in the programme, according to the number they felt they could manage. Although we will aim to post-test all pupils with a pre-test, a sensitivity analysis will look at the impact of this by excluding the pupils who were originally intended to receive the programme but were 'deselected' at random post-randomisation.

Analysis

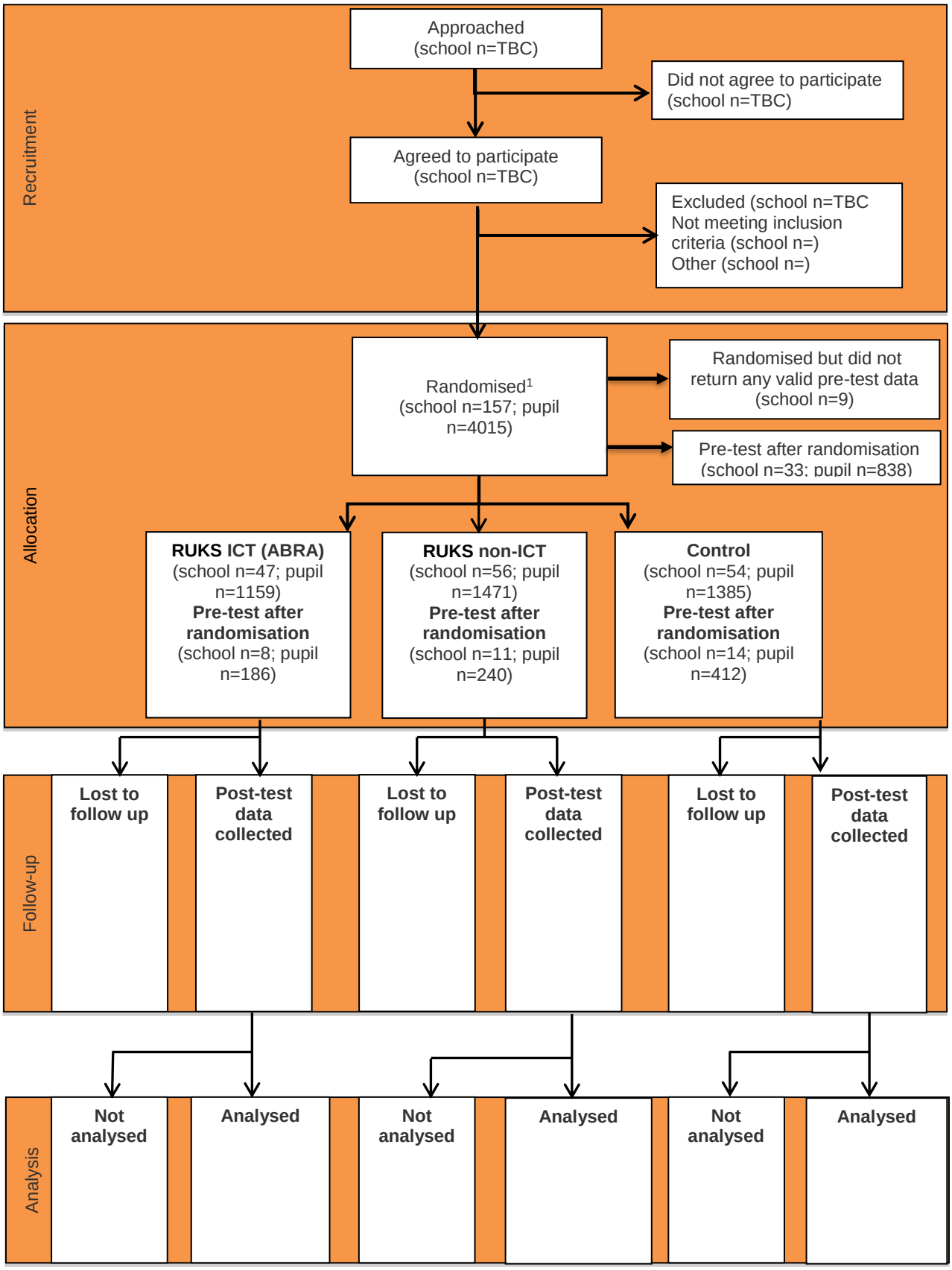
The statistical analysis proposed follows the most recent *revised EEF Statistical Analysis Guidance (2018)* available *here*.

Analysis will be conducted in Stata v15 (or later, to be confirmed in the final report) using the principles of intention to treat (ITT), where data are available, including all schools and pupils in the groups to which they were randomised irrespective of whether or not they actually received the RUKS programme.

Statistical significance will be assessed using two-sided tests at the 5% level. Estimates of effect with 95% confidence intervals (CIs) and p-values will be provided. No formal comparison of baseline data will be undertaken, except to report the differences in PiRA pre-test scores (raw and age-standardised) as a Hedges' g effect size (Hedges, 2007).

A full CONSORT diagram will be produced to show the flow of schools and pupils through the trial (Figure 1).

Figure 1: Participant flow diagram



[1]randomisation took place at the school-level and 166 schools were randomised; however, at the pupil-level, only those with a valid pre-test were considered as randomised. In total, 166 schools were randomised but of these, 9 did not return any valid pre-test data (one of these withdrew before being informed of their allocation). There were therefore 4010 randomised pupils across 157 schools. Of these 157 schools, 33 (385 pupils) conducted the pre-test after being informed of their allocation.

Primary outcome analysis

The raw and age-standardised PiRA post-test scores will be summarised by trial arm, including the number and percentage above the national average (100 for the age-standardised score, relative to PiRA's national standardisation sample). The age-standardised scores are calculated based on the child's age in years and completed months (one month brackets). No further use will be made of raw scores and all proceeding analysis relates to the age-standardised score. The correlation between the pre- and post-test scores will be provided, these will take the form of a raw correlation between pre and post scores, and also we will report the R squared values from the regression models which will represent the proportion of the variability in the outcome variable "explained" by the covariates. Histograms of the pre- and post-test data distributions will be provided.

Two multilevel mixed-effect linear regression models at the pupil-level will be used to compare post-test PiRA age-standardised score between the groups. One model will exclude pupils in the schools randomised to the ICT group, and will be used to investigate the difference between the non-ICT and control groups. The second model will include pupils from all three groups *except* those from the eight schools that were only randomised between the non-ICT and control groups (because they did not have the technology to implement the ICT RUKS programme). This model will be used to obtain the pairwise comparisons between the ICT programme and control groups, and between the ICT programme and the non-ICT programme. Two models are necessary since it is not appropriate to include schools that could never have been allocated to receive the ICT programme in a comparison involving this group. Adjusted differences in scores between pairs of groups will be extracted from the relevant model with a 95% CI and p-value. Both models will be adjusted as follows:

Pupil-level fixed effects:

- Baseline age-standardised PiRA score
- Gender
- FSM (NPD variable EVERFSM_6_P)
- Foundation Stage Profile (NPD variable FSP GLD, defined as whether or not the pupil achieved a good level of development i.e. achieved level of 2 or 3 in each of COM, PHY, PSE, LIT and MAT results.)

School-level fixed effects:

- Allocation (2 or 3 levels, according to model; ICT, non-ICT, control)
- Staff type (3 levels; qualified, non-qualified, both)
- Number of pupils in the Year 1 cohort, as a continuous variable
- Geographical area (5 levels; West Midlands, East Midlands, Newcastle, Teesside, Manchester)

It is customary to adjust analyses for factors used in the stratification/minimisation of the randomisation for a trial; hence, the adjustment here for staff type, geographical area, and number of pupils in the Year 1 cohort. However, since we are adjusting for pupil-level free school meal status, we shall omit school-level percentage of pupils with ever FSM status in the Year 1 cohort as a covariate as these factors are likely to be collinear.

Adjustment will be made for clustering at the school level by including school as a random effect, a standard method for the analysis of cluster trials (Wears, 2002).

Model equation:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 I_{A1i} + \beta_3 I_{A2i} + \beta_4 I_{B1i} + \beta_5 I_{B2i} + \beta_6 y_i + z_i + \beta_8 I_{E1i} + \beta_9 I_{E2i} + \beta_{10} I_{E3i} + \beta_{11} I_{E4i} + u_i + \varepsilon_{ij}$$

Y_{ij} = response of the j^{th} member of the i^{th} cluster (school), i=1, …, m, j=1, …, n_i

m = number of clusters (schools)

n_i = size of i^{th} cluster (school)

x_{ij} = baseline age-standardised PiRA score for j^{th} member of i^{th} cluster (school)

I_{A1i} = 0/1 indicator variable for non-ICT allocation of i^{th} cluster (school)

I_{A2i} = 0/1 indicator variable for ICT allocation of i^{th} cluster (school) (only for model 2)

I_{B1i} = 0/1 indicator variable for qualified staff type of i^{th} cluster (school)

I_{B2i} = 0/1 indicator variable for non-qualified staff type of i^{th} cluster (school)

y_i = number of Year 1 pupils in i^{th} cluster (school)

z_i = proportion of EverFSM pupils in i^{th} cluster (school)

I_{E1i} = 0/1 indicator variable for West Midlands i^{th} cluster (school)

I_{E2i} = 0/1 indicator variable for East Midlands i^{th} cluster (school)

I_{E3i} = 0/1 indicator variable for Newcastle i^{th} cluster (school)

I_{E4i} = 0/1 indicator variable for Teeside i^{th} cluster (school)

β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7, β_8, β_9, β_{10}, β_{11} = fixed effect parameters

u_i = random effect for i^{th} cluster (school)

ε_{ij} = residual error term for j^{th} member of i^{th} cluster (school)

Model assumptions will be checked as follows: the normality of the standardised residuals will be checked using a histogram and qq plot, and the homoscedasticity of the residuals assessed using a scatter plot of fitted values against the residuals. Visual inspection of the plots only will be used (no formal statistical tests). If the model assumptions are in doubt, a sensitivity analysis will be conducted in which transformations of the outcome and/or covariate data will be tried to improve the model fit.

Sensitivity analysis

ISSUES WITH PRE-TEST DATA

We will repeat the primary analyses excluding pupils that completed the pre-test *after* their school was informed of their trial allocation. The PiRA pre-test results for these pupils will be presented alongside those for all other pupils to see if there is any difference. The primary analyses will also be repeated for all pupils using EYFSP result in G09 for Literacy - Reading score (NPD variable FSP_LIT_G09) instead of pre-test PiRA score as the measure of prior attainment.

SELECTION OF PUPILS TO RECEIVE PROGRAMME

As described earlier, following randomisation, some schools asked to deliver the programme to fewer pupils than they had originally stated and allowed YTU to randomly select the pupils to receive the RUKS programme. This happened largely because it became apparent to the schools following the training (and the full details of the RUKS programme becoming apparent), that they would not have capacity to deliver the programme to the original number of pupils. This will likely dilute any treatment effect observed in the primary analyses since it includes pupils that the school could never deliver the programme to. Sensitivity analyses will repeat the primary outcome models but excluding pupils who were pre-tested but then randomly 'deselected' by the YTU to receive the programme immediately following their school attending training.

Secondary outcome analysis

The secondary outcomes of DTWRP, LeST and RAQ will be analysed exactly as described for the primary outcome of PiRA. As these are not assessed at baseline, the PiRA age-standardised score at pre-test will be included as the measure of prior attainment in the models. Sensitivity analyses will also be run for these outcomes excluding the covariate for pre-test PiRA score, and also using instead EYFSP result in G09 for Literacy - Reading score (NPD variable FSP_LIT_G09) as the measure of prior attainment.

To investigate the effects of small group teaching, the primary analyses will be repeated in the subset of pupils who were identified by their schools at baseline as those to be taught in a small group if their school was allocated to teaching as usual. Data provided by the schools on the criteria they used for selecting these pupils and the small group teaching they intended to deliver will be summarised.

Schools will be permitted to group pupils however they see fit for delivery of the programme, and we will aim to record how they do this (by ability, mixed ability, other) via a baseline survey for school staff. Within each programme arm, the number of schools that set the small groups by ability will be presented (compared to mixed ability groups, or another way of composing the groups). Baseline and outcome data will be summarised descriptively, stratified by how the school chose to group the children for the programme. This will be an observational comparison only and so findings will be purely exploratory, but may be used to generate research hypotheses and help steer the direction of future research.

Interim analyses

No interim analyses will be undertaken.

Subgroup analyses

Pupil UPNs, as obtained during the recruitment period, will be used to access additional pupil characteristics from the NPD (e.g. FSM status). The effect of the RUKS programme on pupils who are eligible for FSM will be assessed via the inclusion of FSM status (using the EverFSM indicator EVERFSM_6_P in the NPD) and an interaction term between FSM status and allocation in the primary analysis models. This will be followed by repeating the primary analyses in the subgroup of pupils who have ever been eligible for FSM.

Imbalance at baseline

School and pupil characteristics and measures of prior attainment will be summarised descriptively by randomised group both as randomised and as analysed in the primary analysis models. School data collected at pre-test will include number of pupils in the Year

one cohort, percentage of pupils who are EverFSM in the Year one cohort and percentage of pupils who are EverFSM in the entire school, Pupil-level data includes gender, date of birth (to calculate age in months), EAL, special educational needs (SEN), EverFSM status, current FSM status and EYFSP data. No formal statistical comparisons will be undertaken (Senn, 1994). Continuous measures will be reported as a mean, standard deviation (SD) while categorical data will be reported as a count and percentage. The unadjusted difference between groups on the pre-test PiRA test (raw and age-standardised score) for those analysed in the primary analysis will be reported as an effect size with 95% CI.

Missing data

The amount of missing baseline and outcome data will be summarised, and reasons for missing data explored and provided in the report, where available. A multilevel mixed-effect logistic regression model will be run to assess for statistically significant predictors of missing primary outcome data at the pupil-level, including all available pupil and school-level baseline data as fixed effects, and school as a random effect. Significant predictors and possible mechanisms for the missing data will be discussed in the report.

If more than 5% of randomised pupils are excluded from the primary analysis due to missing data, then the impact of missing data on the primary analysis will be additionally assessed using multiple imputation by chained equations, predicted by pre-test PiRA age-standardised score, school, allocation and any variables found to be significant in the 'drop-out' model described above.

A 'burn–in' of 150 will be used and 30 imputed datasets will be created. The primary analysis will then be rerun within the imputed datasets and Rubin's rules will be used to combine the multiply imputed estimates.

Compliance

Attendance of school staff members at the training events will be reported. Programme schools will be encouraged to use the RUKS programme during the course of one academic school year (2018/2019) for a minimum of 20 weeks, but will be able to continue beyond 20 weeks at their choice. Based on the evidence from the previous efficacy trial, schools will be instructed to group pupils in Year 1 into small groups of 3 to 4 pupils and deliver the RUKS programme in four 15 minute sessions per week, supported by a member of school staff. This is the same for both the ICT and non-ICT groups; the RUKS programme is delivered in sessions, the only difference being the nature of delivery i.e. whether using the on-line platform or paper-based material.

Schools will be asked to keep registers to indicate when and which pupils partake in programme sessions in the ICT and non ICT groups. Data from the registers will be entered by the Delivery Team into Excel spreadsheets and sent to the YTU via DropOff. These data will be summarised. A Complier Average Causal Effect (CACE) analysis for the primary outcome will be carried out to assess the effect of the RUKS programme in the compliers. Compliance will be defined at the pupil-level in three ways as follows:

- Minimal compliance – completed at least four sessions of the programme (Y/N)
- Full compliance – completed 80% (n=64) of the planned 80 sessions (Y/N)
- Number of sessions completed (continuous variable)

A two-stage least squares instrumental variable (IV) approach will be used, using random group allocation as the IV (Dunn et al., 2005).

Intra-cluster correlations (ICCs)

The school-level intracluster correlation coefficient (ICC) for the post-test outcomes will be extracted from each multilevel analysis model, with the 95% CI. The ICC associated with school for the pre-test scores will also be presented with a 95% CI.

Effect size calculation

Effect sizes will be calculated by dividing the adjusted mean difference between the RUKS programme and control group by the pooled variance obtained from the unconditional model.

$$ES = \frac{(\bar{Y}_I - \bar{Y}_C)_{adjusted}}{\sqrt{s^*}}$$

Where,

$(\bar{Y}_I - \bar{Y}_C)_{adjusted}$ denotes the difference in means between the trial arms obtained from the adjusted analysis mixed model, and

s^* denotes the pooled unconditional variance from a mixed model run with only adjustment for trial arm and clustering at school-level. The pooled variance is obtained by the sum of the between- and within-cluster variance.

A 95% CI for the effect size will be calculated by dividing the 95% confidence limits for the adjusted mean difference by the same standard deviation. All parameters used in these calculations will be provided in the final report.

References

ABRAMI, P. C., SAVAGE, R. S., DELEVEAUX, G., WADE, A., MEYER, E. & LEBEL, C. 2010. The learning toolkit: The design, development, testing and dissemination of evidence-based educational software 1. *Design and implementation of educational games: Theoretical and practical perspectives.* IGI Global.

COMASKEY, E. M., SAVAGE, R. S. & ABRAMI, P. 2009. A randomised efficacy study of Web-based synthetic and analytic programmes among disadvantaged urban Kindergarten children. *Journal of Research in Reading,* 32**,** 92-108.

DEPARTMENT OF EDUCATION 2017. National Curriculum Assessment at Key Stage 2 in England (interim). *Department for Education Crown Copyright* *https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/624576/SFR30_2017_Text.pdf Last accessed 15.01.2018*.

DUNN, G., MARACY, M. & TOMENSON, B. 2005. Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. *Statistical methods in medical research,* 14**,** 369-395.

FREIDLIN, B., L KORN, E., GRAY, R. & MARTIN, A. 2008. *Multi-Arm Clinical Trials of New Agents: Some Design Considerations*.

JOHNSON, H., MCNALLY, S., ROLFE, H., RUIZ-VALENZUELA, J., SAVAGE, R., VOUSDEN, J. & WOOD, C. 2019. Teaching assistants, computers and classroom management. *Labour Economics,* 58**,** 21-36.

MCNALLY, S., RUIZ-VALENZUELA, J. & ROLFE, H. 2016. ABRA: Online Reading Support. Evaluation Report and Executive Summary. *Education Endowment Foundation*.

SAGHAEI, M. & SAGHAEI, S. 2011. Implementation of an open-source customizable minimization program for allocation of patients to parallel groups in clinical trials. *Journal of Biomedical Science and Engineering,* Vol.04No.11**,** 6.

SAVAGE, R., ABRAMI, P. C., PIQUETTE, N., WOOD, E., DELEVEAUX, G., SANGHERA-SIDHU, S. & BURGOS, G. 2013. A (Pan-Canadian) cluster randomized control effectiveness trial of the ABRACADABRA web-based literacy program. *Journal of Educational Psychology,* 105**,** 310.

SAVAGE, R. S., ABRAMI, P., HIPPS, G. & DEAULT, L. 2009. A randomized controlled trial study of the ABRACADABRA reading intervention program in grade 1. *Journal of Educational Psychology,* 101**,** 590.
SENN, S. 1994. Testing for baseline balance in clinical trials. *Statistics in medicine,* 13**,** 1715-1726.
SHARPLES, J., WEBSTER, R. & BLATCHFORD, P. 2015. Making best use of teaching assistants: Guidance report.
TORGERSON, C., BROOKS, G., GASCOINE, L. & HIGGINS, S. 2018. Phonics: reading policy and the evidence of effectiveness from a systematic 'tertiary'review. *Research Papers in Education***,** 1-31.
WASON, J. M. S., STECHER, L. & MANDER, A. P. 2014. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials,* 15**,** 364-364.
WEARS, R. L. 2002. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. *Acad Emerg Med,* 9**,** 330-41.